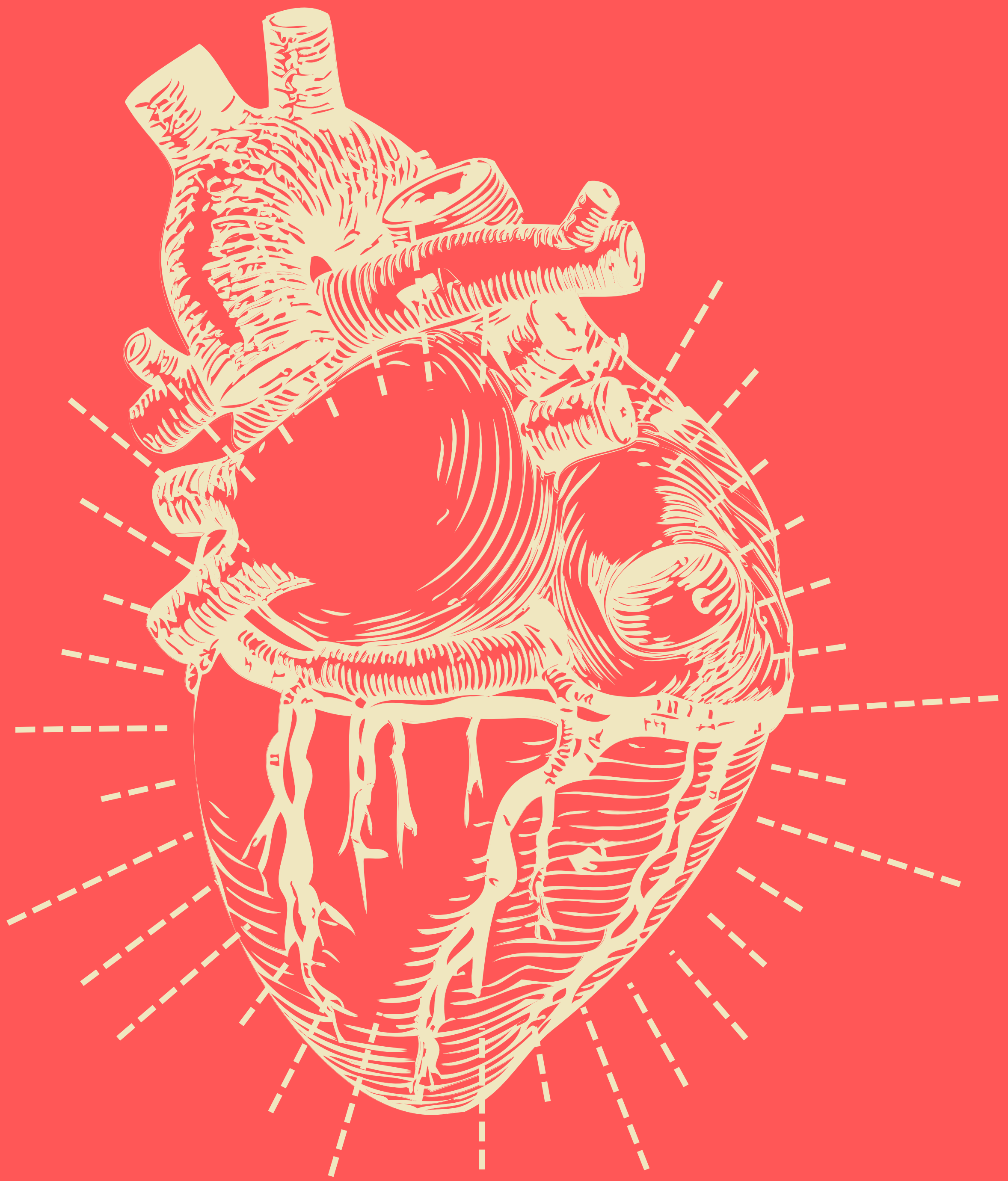
Classifying if a Patient Has Heart Disease

BY TEAM 3:
AKASH RUDRA, ANDREW TRAPP, MAHIMA MASETTY, TAYLOR LEGGETT, VICTOR FLORIANO



Using Machine Learning To Detect Heart Disease

What is the problem?

1 out of every 4 deaths in the United States is from a form of preventable heart disease

90% of heart disease is preventable through healthy diet, regular exercise and not smoking

How are we going to solve it?

Machine Learning can be used to deduce whether a patient is at risk of developing heart disease based on their medical statistics. Based on the variables contributing their likeliness to get heart disease, care providers can suggest aspirin, smoking cessation, blood pressure control, and cholesterol management accordingly.

Describing Our Dataset

ROWS

Our Dataset was created by combining 5 independent datasets and it is available on Kaggle.

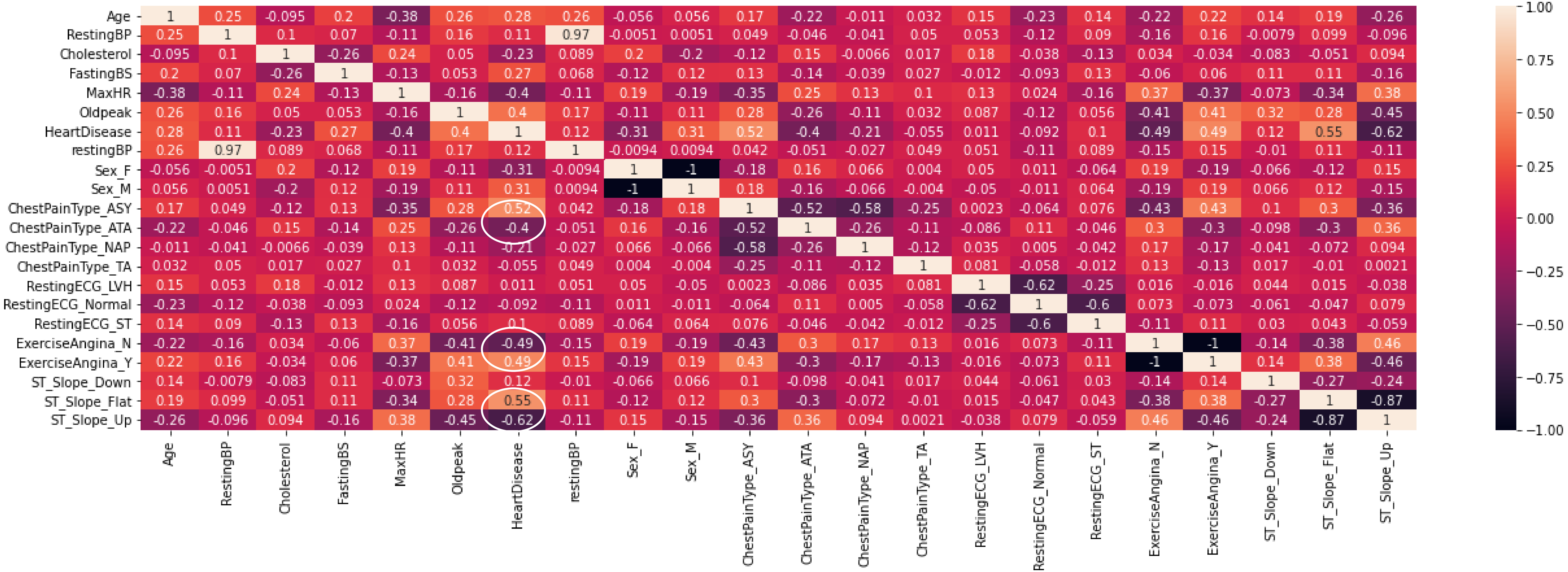
- Cleveland: 303 observations
- Hungarian: 294 observations
- Switzerland: 123 observations
- Long Beach VA: 200 observations
- Stalog (Heart) Data Set: 270 observations

Final number of rows: 918
Duplicated rows: 272
Total: 1190 observations

COLUMNS

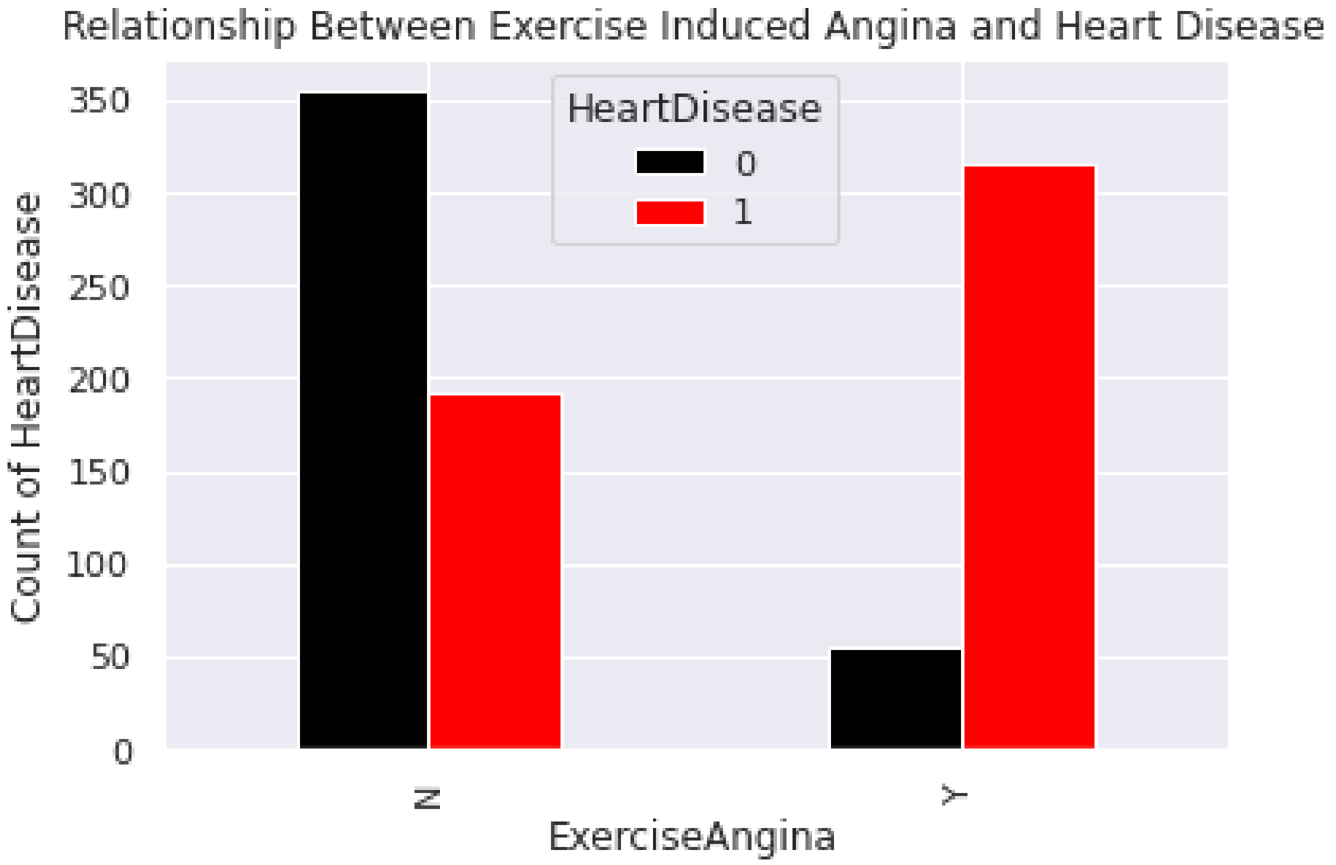
15 Predictors
6 Numerical | 9 Categorical

Finding Correlations Between Variables To Inform Our Initial Hypothesis

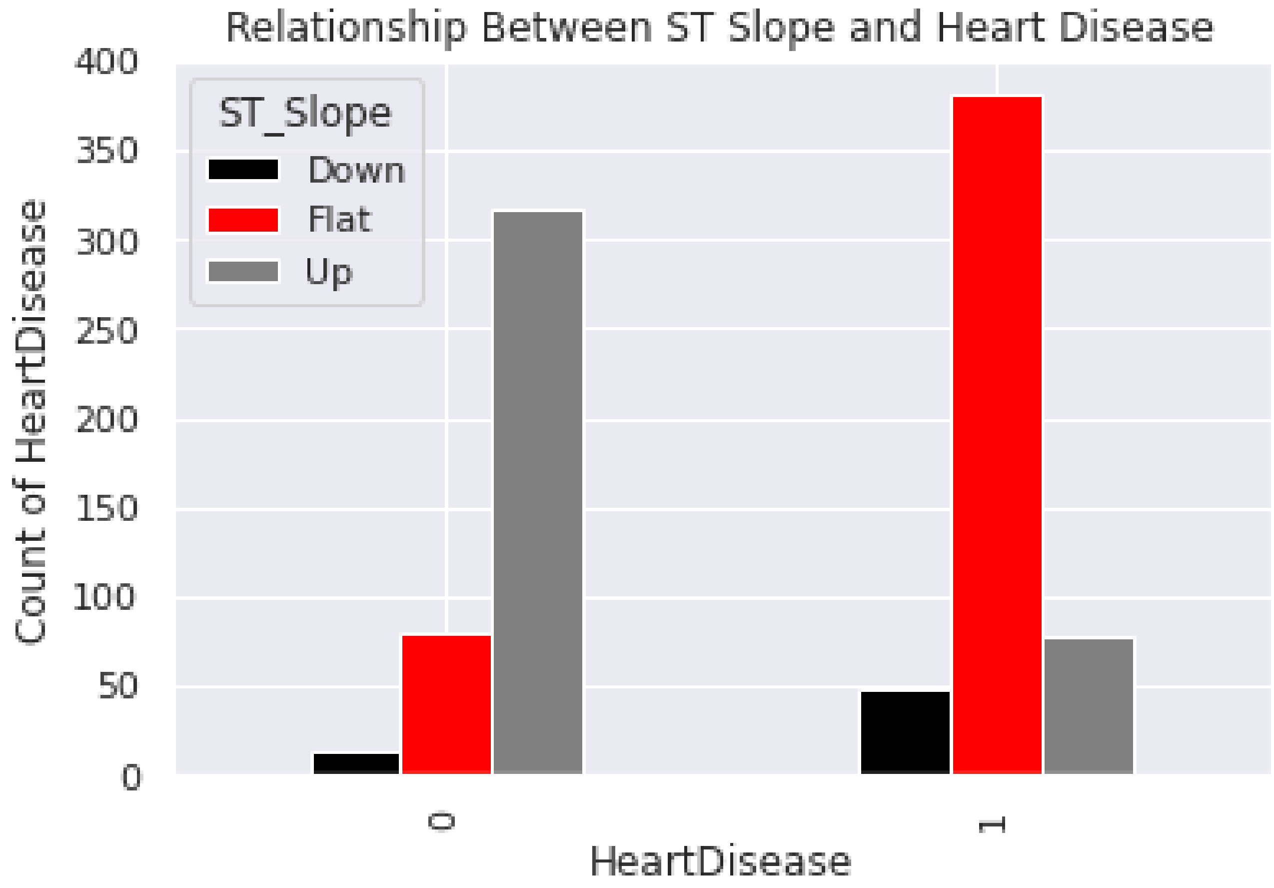


Investigating the Correlations Further

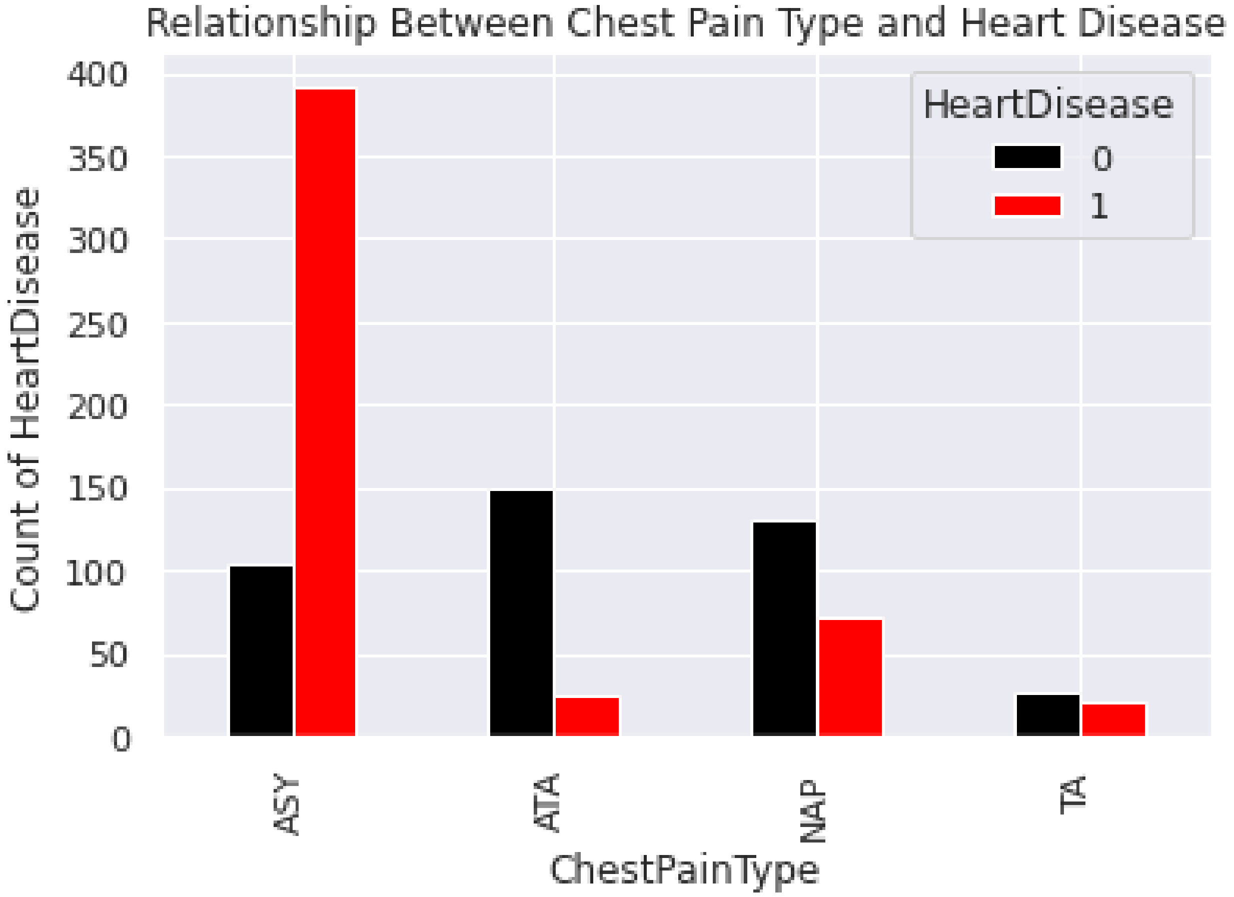
1.



2.



3.



1. If the patient suffers from exercise induced angina, they are more likely to have heart disease
2. If the ST slope of a patient is flat, they are more likely to have heart disease while if the ST slope is up, they are less likely to have heart disease
3. If a patient has ASY type of chest pain, they are very likely to have heart disease

Pre-Processing To Improve Data Usability

Remove Duplicate Instances

- Originally 1190 instances, but 272 of them are duplicates
- Drop to remove model bias
- Number of instances in the end: 918

Dummy Variables for Categorical Predictors

- Categorical Predictors: 'Sex', 'ChestPainType', 'RestingECG', 'ExerciseAngina', 'ST_Slope'
- Dropped first class of each

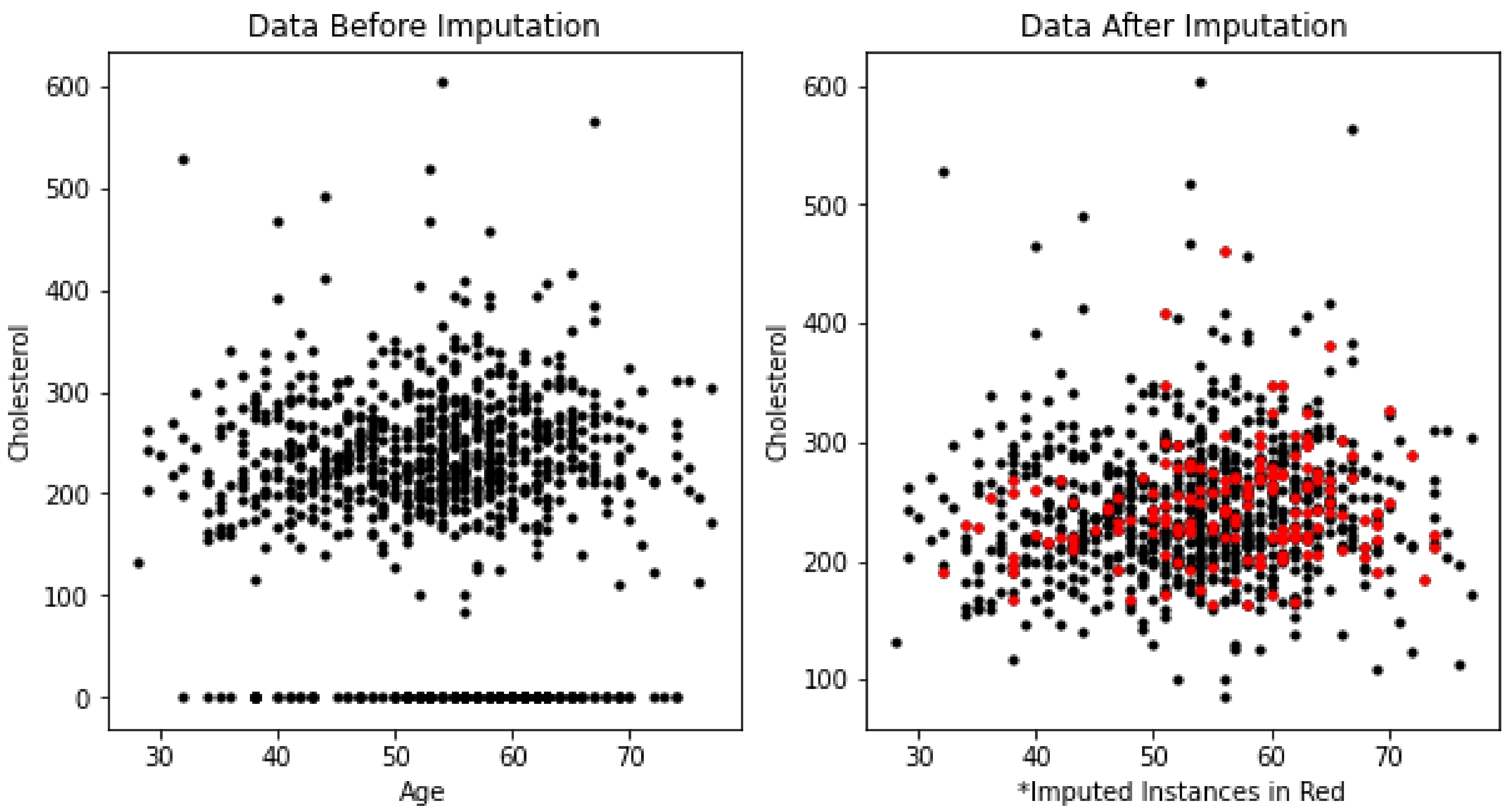
Outliers

- Used IQR to check for outliers
- Predictive power got worse when the outlier instances were removed so they were not removed

Imputation

- One instance missing in resting_bp value which we imputed using mean
- Address missing Cholesterol value by comparing different imputation methods

Imputation Used To Fill Missing Cholesterol Values



PROBLEM

There are about ~200 instances in which the "Cholesterol" variable is inputted as 0 and we decided to impute these values rather than drop them.

SOLUTION

Tested a variety of different univariate (mean, median, mode) and multivariate (kNN, MICE) imputation methods. Chose kNN imputation, as it produced the best F1 score using logistic regression, of the 5 imputation methods.

Exploring Interaction Terms for Feature Engineering

GOAL

Testing interaction terms to identify if the marginal effect of X1 on Heart Disease was dependent on X2

METHODOLOGY USED

- Selected a group of predictors to test and added an interaction term
  - **Age and MaxHR –** age*maxHR
  - **Age and RestingBP** – age*restingBP
  - **Gender and MaxHR** – gender*maxHR
  - **Oldpeak and ST_Slope_UP** – oldpeak*ST_Slope_Up

2. Ran a regression of Heart Disease by one pair of predictors (baseline), then another with one pair of predictors and its interaction term to evaluate p-values

RESULT

Interaction term for **oldpeak*slope_up** was the only one significant at the 1% level and added to our dataset

Using F1 Score To Measure Our Models

- Dealing with medical data we **could not use accuracy** because for us a False Negative is a lot more costly than a False Positive
- **Recall(True Positive Rate) would be a good metric,** but a model could have 100% recall by predicting Heart Disease for every instance
- **We decided to go with F1** which is the harmonic mean of recall and precision

$$F1\,score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Creating a Baseline Model as a Standard for Our Conclusions

WHAT IS IT?

An 'oversimplified model' to act as a reference for us to contextualize the results of our other models to see which had the best predictive power.

HOW DID WE CREATE IT?

- GroupBy 'Outcome' divided by total instances
- Find the maximum of the two fractions

WHAT IS THE OUTCOME?

If we predict everyone in our dataset has Heart Disease, we will be correct 55.34% of the time, meaning we will be wrong 44.66% of the time.

Our Decision Tree Has High Variance and High Interpretability

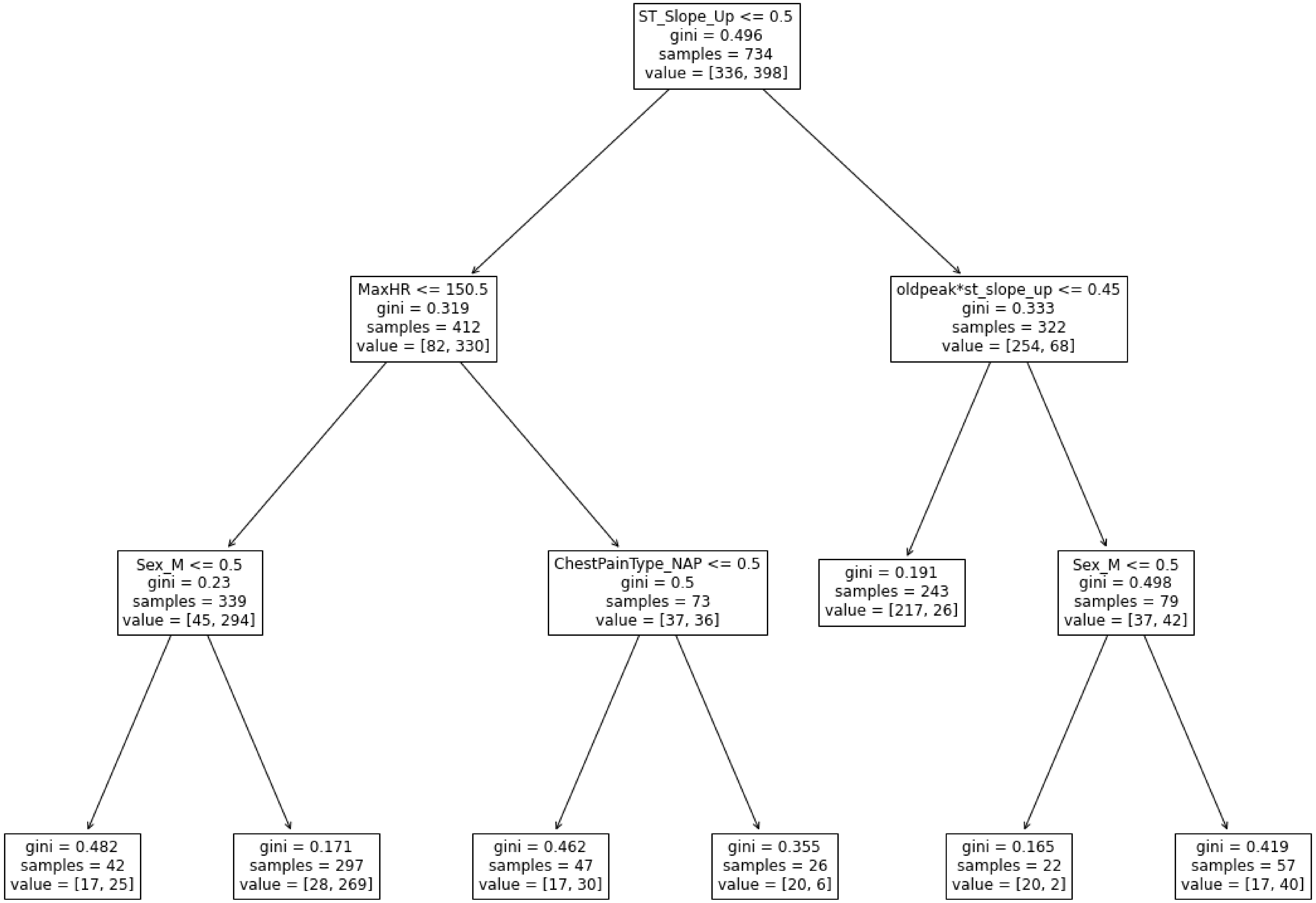
Tuning Parameter: ccp_alpha | Method: Cross-Validation (cv=5)

DECISION TREE	
CCP_ALPHA	0
F1 SCORE	0.788

DECISION TREE TUNED	
CCP_ALPHA (BEST)	0.007
F1 SCORE	0.877

Issue: Due to high variance in Decision Trees, we observed that the scores metrics would vary significantly with changes in the Train/Test split

Our Decision Tree Has High Variance and High Interpretability



Our Random Forest Has Lower Variance but Low Interpretability

Tuning Parameters: ccp_alpha, n_estimators | Method: GridSearch CV

RANDOM FOREST	
N_ESTIMATORS	100
CCP_ALPHA	0
F1 SCORE	0.880

RANDOM FOREST TUNED	
N_ESTIMATORS	100
CCP_ALPHA	0.001
F1 SCORE	0.898

Improvement over DT: Better score metrics and lower variance

Visualizing Our Random Forest with a Reborn Decision Tree

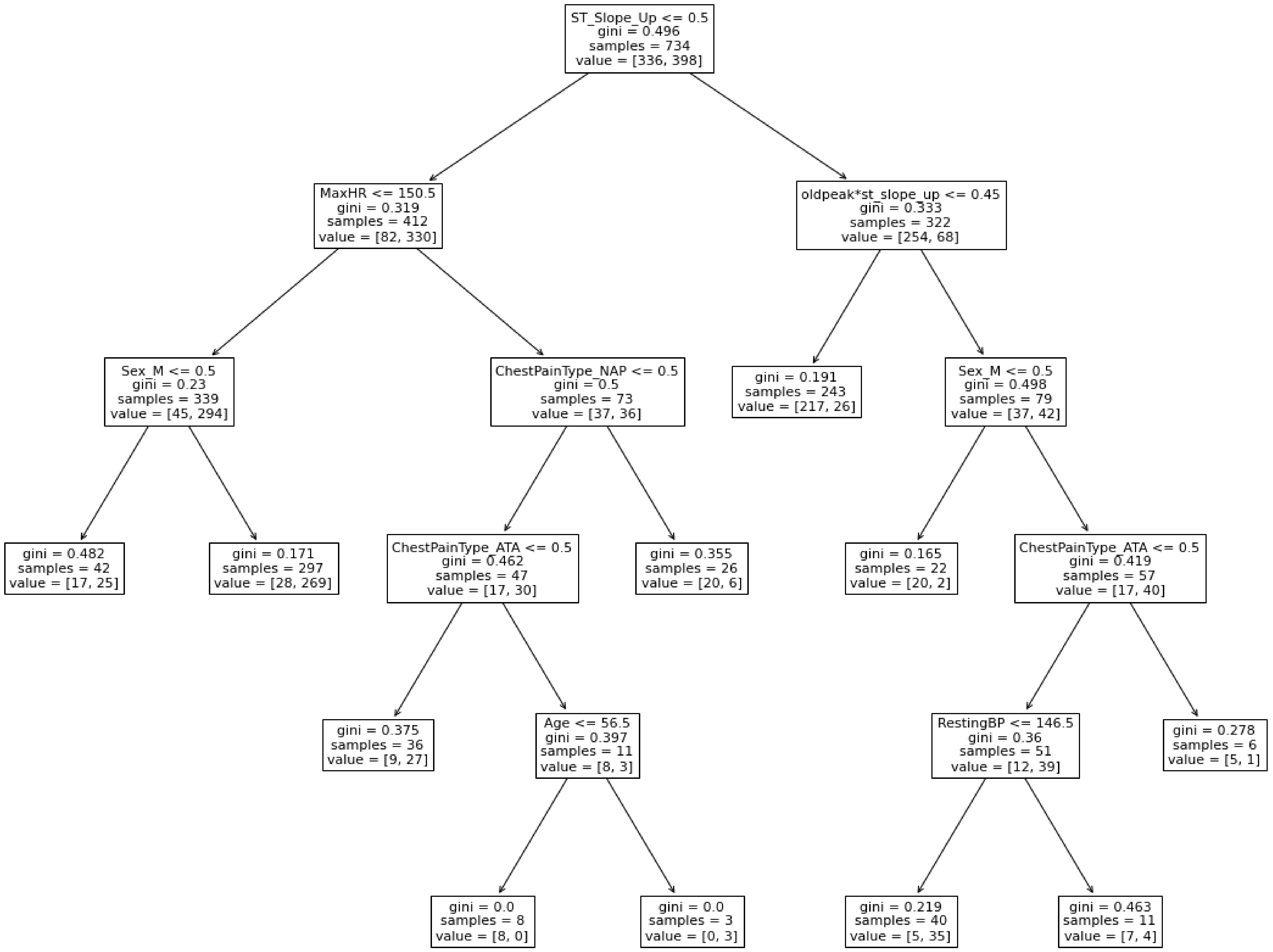
METHODOLOGY USED

1. Stored our Random Forest predictors in a new variable: **RF_predictions**
2. Instantiated a new Decision Tree with **RF_predictions** as our **y_train**
3. Compared the predictions made by the Random Forest with our Reborn Tree

RESULTS

A new Decision Tree that makes the same predictions as our Tuned Random Forest with 100% accuracy over the training data

Visualizing Our RF with a Reborn DT



Parameter Tuning and Performance for KNN Model

Tuning Parameters: K (neighbors), Weights (uniform or distance)

KNN TUNED	
K (BEST)	25
WEIGHTS (BEST)	DISTANCE
F1 SCORE	0.927

Method: Standardized data, GridSearch CV

Identifying the Best Parameters and Threshold for our Logistic Regression

LOGISTIC REGRESSION WITH PARAMETERS	
THRESHOLD	.5
F1 SCORE	0.900

Through tuning the parameters of the model, we found the new model matches the baseline model

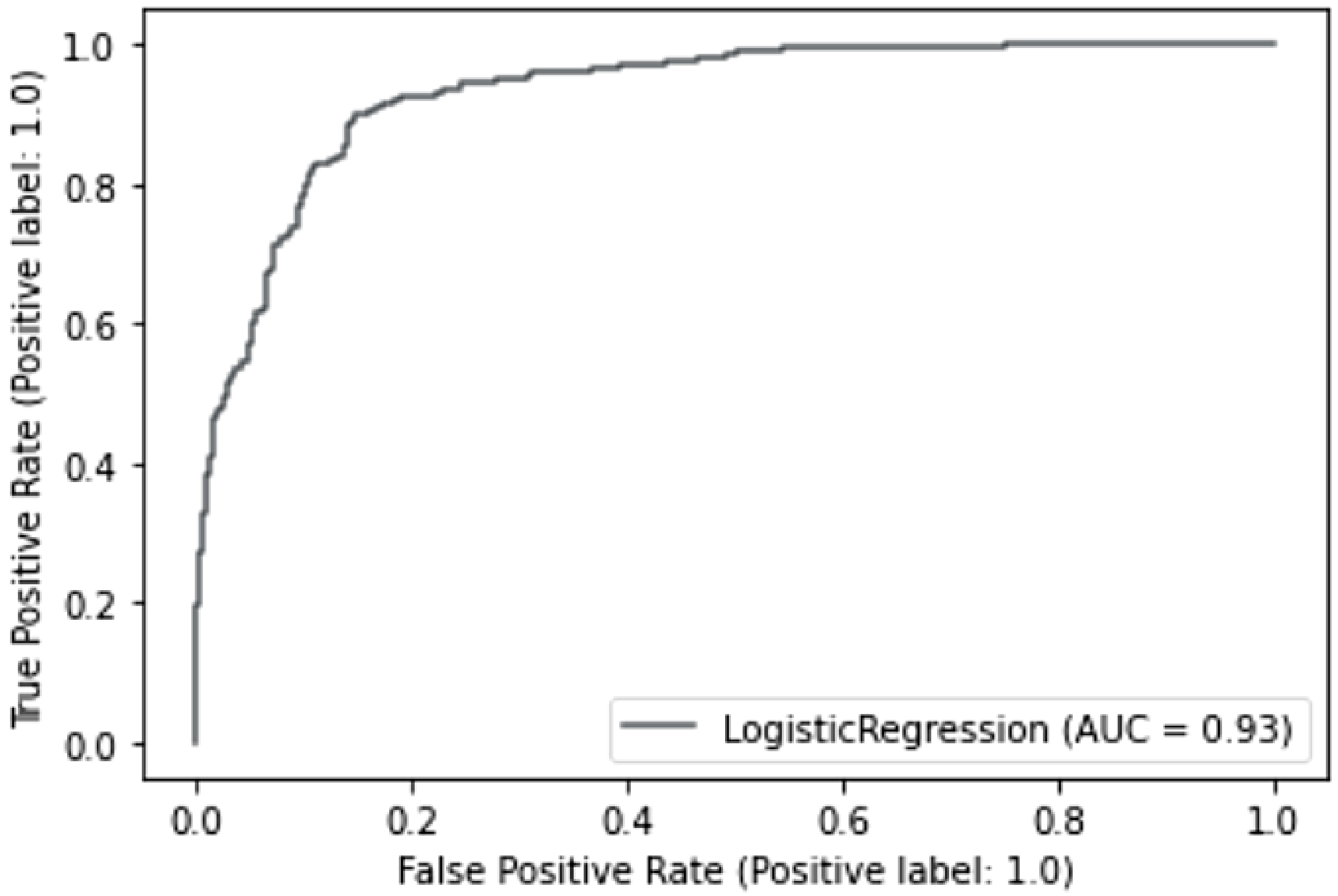
- The best penalty parameter was "none"
- C= .00001
  - This parameter is not used because the penalty parameter is "none"
- Best threshold found was .5
  - Which is the same as the default parameter in the baseline model

Tuning Our Model To Find the Best Threshold

ROC Curve:

Best Threshold = 0.5

Same as the default threshold used in the baseline model



Evaluating Performances To Find the Best Model

MODEL TYPE	F1 SCORES
BASELINE	0.553
LOGISTIC	0.911
RIDGE	0.864
DECISION TREE	0.877
RANDOM FOREST	0.898
KNN	0.927

Random Forest Metrics

Accuracy: 0.88

Recall: 0.882

IDENTIFYING THE OBSTACLES WE FACED AND HOW WE CAN ENHANCE OUR FINDINGS

OBSTACLES	IMPROVEMENTS

• Increase the amount of data
• Balanced dataset
    ○ Specifically the number of men vs women
• More specific information on individual patients
    ○ Location, past habits (smoking), etc.
	• If we had more information on the patients, we would have been able to do more in depth analysis
	    ○ Does location have an impact on heart disease?
	    ○ Does smoking increase chances of heart disease?

The Best Predictors

FACTORS THAT CONTRIBUTE THE MOST TO HEART DISEASE

- ST slope: ST_Slope_Up
- Max Heart Rate
- Sex: Male
- Chest Pain : NAP
- Old Peak * ST_Slope_Up

Looking Forward

OUR MODEL ENCOURAGES TAKING PREVENTATIVE MEASURES

- Keep an eye on the patients that fall into the categories above
- Perform frequent procedures to check condition of their heart

Thank you
Questions?